COVID-19 Data Project

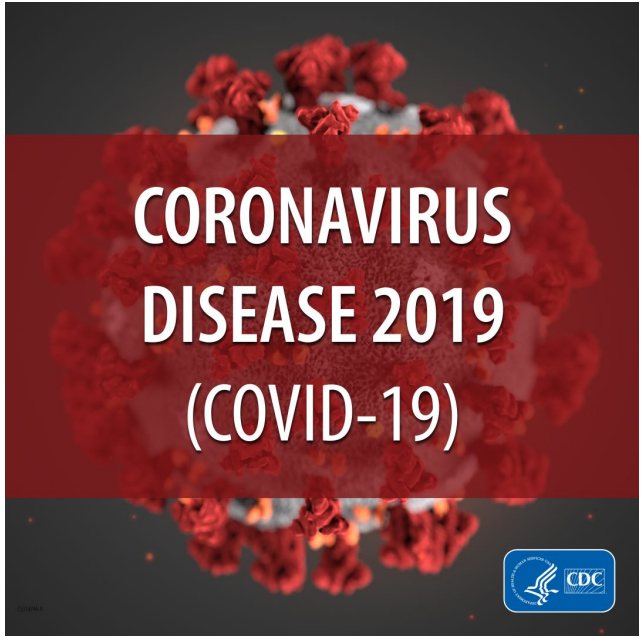
Patrick Holder

The Problem

- A coronavirus identified in 2019, SARS-Cov-2, has caused a pandemic of respiratory illness, called Covid-19.
- Early on in the virus's history, it was clear that certain conditions made the disease more dangerous.
- The risk of developing dangerous symptoms of COVID-19 has been linked to heart or lung conditions, weakened immune systems, obesity, diabetes, asthma, and others.
- Medical supplies needed to treat these dangerous symptoms of COVID-19 were often in short supply during the worst periods of the pandemic.

A Possible Solution

- Data analysis of patient information will allow your company to relocate more supplies in areas where patients are at a greater risk of serious symptoms or death.
- We will also provide a machine learning model to better identify individuals who are at a higher risk of death when they are diagnosed with COVID-19.



The Data

- A dataset consisting of 21 unique features and 1,048,576 unique patients was acquired from the Mexican government.
- In the data a 1 means a “yes and a 2 means a “no”

- Existing Conditions Included:
  - Pneumonia
  - Pregnancy
  - Diabetes
  - Chronic Obstructive Pulmonary Disease
  - Asthma
  - Immunosuppressed
  - Hypertension
  - Cardiovascular disease
  - Chronic renal disease
  - Obesity
  - Chronic tobacco use
  - Other diseases
- Other Data Included
  - Age/Sex
  - Classification
  - Hospitalized
  - Medical Unit
  - Intubated
  - ICU
  - Date of Death

The Process

Some Challenges:

- Some of the data was not used for the analysis and was dropped from the data
  - INTUBATED, ICU, MEDICAL_UNIT, USMER
- Some of the data needed to be changed
  - Classification - describes whether or not the patient was diagnosed with covid
  - DATE_DIED - This either has a date or if the patient survived it is indicated with “9999-99-99”

Classification

- Values 1-3 mean that the patient was diagnosed with covid in different degrees. 4 or higher means that the patient is not a carrier of covid or that the test was inconclusive.
- The Process:
  - We first changed this to a boolean feature
  - We then dropped any data coming from patients who were not diagnosed with COVID-19

DATE_DIED

- We wanted to change this data to a boolean feature
- The Process:
  - We generated a new DataFrame with 2 columns and one entry
  - We then did an outer join with the new table and the original data.
  - We then changed every null value to 1.

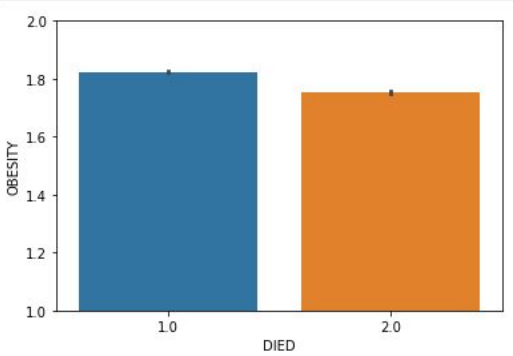
DATE_DIED	DIED
999-99-99	2

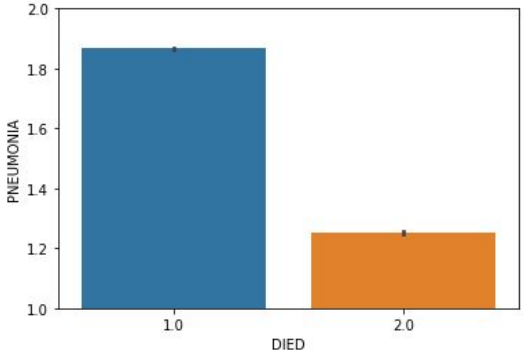
Some Analysis:

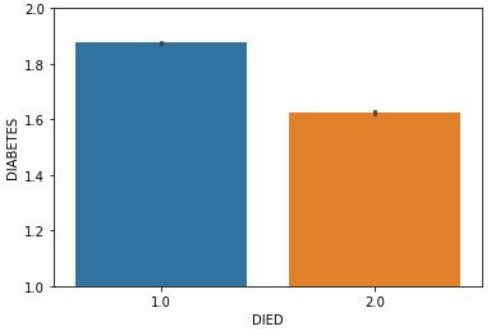
DIED	SEX	PATIENT_TYPE	PNEUMONIA	AGE	PREGNANT	DIABETES	COPD	ASTHMA	INMSUPR	HIPERTENSION	OTHER_DISEASE	CARDIOVASCULAR	OBESITY	RENAL_CHRONIC	TOBACCO
1.0	1.649644	1.906765	1.250750	61.549869	1.998838	1.623669	1.952980	1.979873	1.973894	1.569771	1.950244	1.947170	1.754573	1.933883	1.918966
2.0	1.515409	1.182953	1.866454	42.537596	1.992006	1.876075	1.989363	1.972350	1.990054	1.841281	1.978255	1.983312	1.823667	1.986990	1.927816

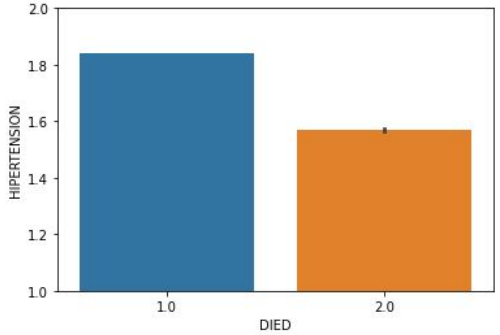
PATIENT_TYPE	SEX	PNEUMONIA	AGE	PREGNANT	DIABETES	COPD	ASTHMA	INMSUPR	HIPERTENSION	OTHER_DISEASE	CARDIOVASCULAR	OBESITY	RENAL_CHRONIC	TOBACCO	DIED
1	1.501069	1.956447	40.937831	1.992641	1.902787	1.991949	1.971785	1.992340	1.866106	1.981409	1.986425	1.832388	1.990864	1.928358	1.982175
2	1.617123	1.338329	55.847398	1.993714	1.685458	1.965099	1.977444	1.976387	1.646175	1.956618	1.957830	1.767904	1.951325	1.922136	1.559211

Some Observations:



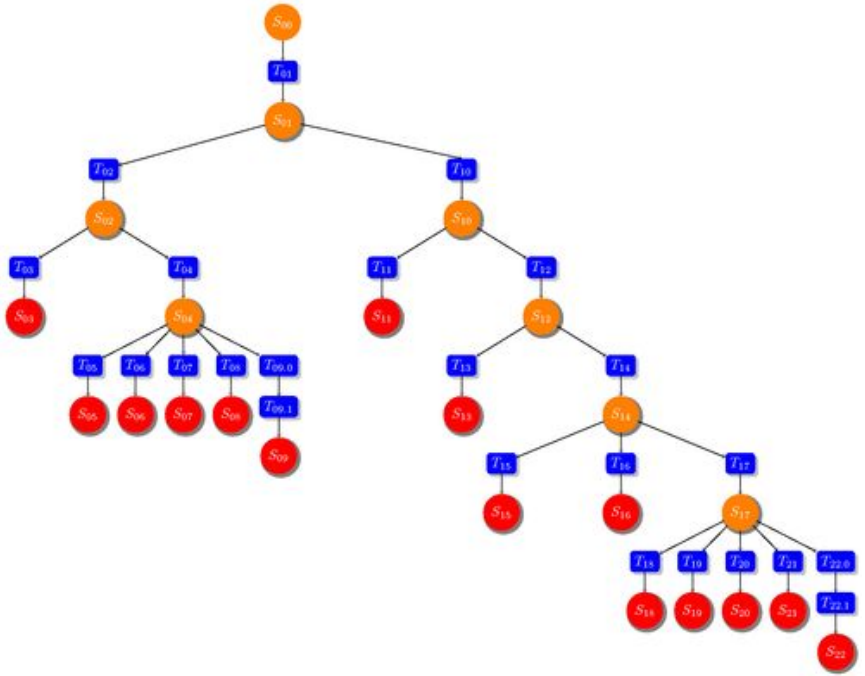




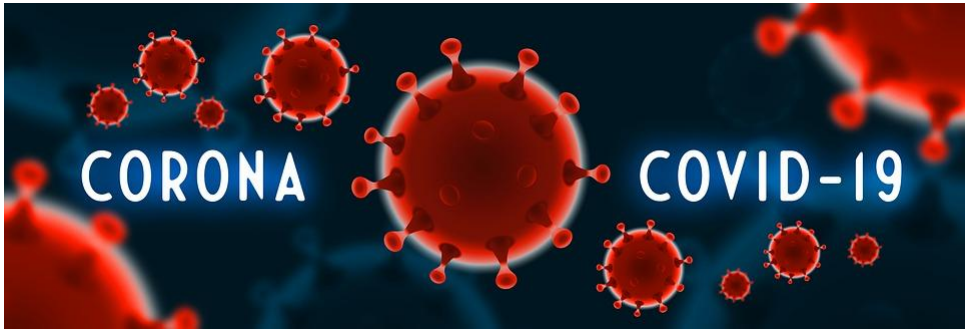


More Observations

- We trained two different computer learning algorithms on this data:
  - Logistic Regression
  - Decision Tree
- Decision Tree generally performed better



Recommendations:



- Send more medical supplies to areas with higher occurrences of Diabetes, Pneumonia, Hypertension, and Obesity.
- Use the machine learning model to find patients with a higher change of developing fatal symptoms, and provide early intervention.

Future Work

- Data outside of Mexico needs to be included to improve the model
- Different algorithms may provide more accurate results
- More research needs to be conducted.